Advance Access publication October 24, 2008

Political Analysis (2008) 16:464–477
doi:10.1093/pan/mpn008

Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors Over Time and Space

Stephen M. Shellman
The Institute for the Theory and Practice of International Relations, The College of William and Mary, P.O. Box 8795, Williamsburg, VA 23186
e-mail: smshel@wm.edu

This article describes a new machine-coded event data set specifically designed to study the spatially, temporally, and tactically disaggregated actions of multiple state and nonstate actors in a systematic fashion. The project develops an extensive set of dictionaries for multiple actors and employs a new coding scheme to organize information on such actors and their behavior. The author describes the machine content-analysis methods used to collect the data and the newly developed coding scheme.

1 Introduction

With too few exceptions, country-year national attribute studies dominate the quantitative study of civil conflict. Interactions among multiple actors engaged in civil conflicts, most notably rebels and governments, in contrast, take place on a day-to-day basis across different parts of a territory. Recently, scholars have pointed out that intrastate conflict is not always characterized as a single conflict between one dissident group and a government, but instead governments often face multiple challengers fighting for the same cause and/or very different causes. On occasion, conflicts involve infighting among members or branches of the government (e.g., military coups in Nigeria) and some conflicts yield dissident group splits (e.g., the Moro Islamic Liberation Front [MILF] emerged out of the Moro National Liberation Front). In other cases, multiple groups may interact with each other and even form alliances or coalitions (e.g., the Coalition Government of Democratic Kampuchea—comprised by the Khmer Rouge, *Front Uni National pour un Cambodge Indépendant, Neutre, Pacifique, et Coopérati* [FUNCINPEC], and Khmer People's National Liberation Front [KPLNF]). In sum, intrastate conflict is comprised of many different players with different motivations, who make a variety of decisions as to how to behave in both the short and long run, in different spatial locales.

Author's note: I would like to thank Philip Schrodt for his help and guidance with this project over the last few years. He did everything from answering numerous e-mails to fixing small programming errors in a moment's notice. I could not have completed this project without his time, patience, and support. I would also like to thank Brandon M. Stewart for his valuable research assistance, ideas, and strong work ethic over the years as he worked on this project. Finally, I would like to thank the anonymous reviewers and guest editors for helpful comments on earlier drafts of this essay. *Conflict of interest statement*: None declared.

© The Author 2008. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Yet, to date, most studies take the monadic aggregate approach to studying civil conflict. The article outlines the theoretical reasons behind creating a disaggregated intrastate conflict event data set to study behavioral interactions among multiple actors fighting each other in different locales on a day-to-day basis using automated natural language processing techniques that save time and money while increasing validity and reliability.

The article begins with a brief discussion of the current state of the civil conflict literature that serves as a springboard for describing the new machine-coded data set. Next, the article outlines the advantages of machine-coding events and the innovations of Project Civil Strife (PCS). Subsequently, the article reports summary statistics for a few variables that illustrate some of the project's contributions. The article closes by providing a glimpse into the types of models we can estimate using the new data and discussing the future of automated event data coding.

2 The Civil Conflict Literature

Most of the classic quantitative intranational conflict research (i.e., civil war, contentious politics, repression and dissent, terrorism, etc.) examines the relationship between country attributes and conflict variables at the country-year level of analysis. Such studies analyze the presence, amount, or level of conflict a country experiences as a function of a country's economy, society, or regime type (e.g., Hegre et al. 2001, Sambanis 2001, Fearon and Laitin 2003, Collier and Hoefler 2004). Alternatively, the conflict processes literature favors an actor-based behavioral approach to the study of civil conflict rather than a systems or structural approach. Such an approach focuses attention on the strategic choices that actors make when engaged in a political struggle. Studies focusing on actors argue that dissidents respond to states and states respond to dissidents (e.g., see Tilly 1985; Lichbach 1987; Davenport 1995; Moore 2000; Shellman 2006a, 2006b; Thyne 2006). Moreover, the population interacts with governments and dissidents and can play important roles in conflict dynamics. Although we have learned from the structural approach that characteristics of the state such as regime type, the economy, terrain, capabilities, and demographics like population and ethnicity are correlated with the level of political conflict we observe *across countries*, the approach has not taught us much about conflict processes as they unfold *over time within specific countries.*

Large-N country-year studies also suffer from temporal and spatial over-aggregation. Goldstein and Pevehouse (1997) argue that ''high levels of aggregation (such as annual data) tend to swallow up important interaction effects.'' Yearly aggregations tend to hide the actions and reactions of actors who respond to one another in much smaller units of time such as weekly intervals (Shellman 2004). Moreover, civil conflicts rarely span an entire country's territory. Rather, they are often confined to subnational regions based on certain geographic features that generate conditions favorable for conflict. If we wish to investigate theories that have a geographic element, we should abandon the country level of analysis in favor of a disaggregated spatial approach (Rød and Buhaug 2007).

Finally, we know that dissidents and governments often employ a range of tactics over the course of a political confrontation. Yet, most of the literature focuses on a particularistic conceptualization of behavior (e.g., terror, war, repression, human rights abuses, protests, etc.). By adopting a particularistic conceptualization of conflict behavior, we ignore the full underlying dimension of state and nonstate actors' behavior (See Moore 2006, 1).

The PCS data sets, described below, were developed to study the strategic behavior of governments, dissident groups, and members of the population by collecting information on a range of tactics such that we can theorize and empirically model a more holistic

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

conceptualization of political behavior. PCS directly confronts actor, spatial, behavioral, and temporal biases by collecting information on multiple actors' behavioral interactions each day in various geographic locations.

3 The Benefits of Automated Natural Language Event Coding

One fruitful way to study multi-actor conflicts unfolding over time on a day-to-day basis is through the collection and analysis of disaggregated event data. Event data are "day-by-day coded accounts of who did what to whom as reported in the open press," which "offer the most detailed record of interactions between and among actors" (Goldstein 1992, 369). Most basic event data sets code the (1) actor taking the action, (2) the target receiving the action, (3) the action itself (the event), and (4) the date of the action/event (usually the day each event takes place). Many hand-coded projects also code the location of the event and/or the number of casualties associated with the event. Some example events coded in political violence data sets include armed attacks/conflict, nonviolent protests, negative statements, positive statements, low-level agreements between actors (e.g., ceasefires), and high-level agreements between actors (e.g., regional territorial autonomy).

Until recently, the collection of event data was cost prohibitive for most researchers. Historically, event data were coded manually, leading to problems such as low inter-coder reliability and a lack of coder attention to detail over time as they spend countless hours reading documents, identifying events from text, and classifying them into different categories.

In the early 1990s, the Kansas Event Data System (KEDS) demonstrated that the collection of event data could be automated (see Schrodt and Gerner 1994; Schrodt, Davis, and Weddle 1994). With automated coding, the coding rules are transparent, the data are easily and quickly reproducible, the data can be regenerated using alternative coding schemes, and the data are unaffected by individual coders' biases, as well as reducing the time required for coding from hundreds of hours of human labor to mere minutes once the input texts have been formatted and coding dictionaries prepared. This has radically changed the information that is available to conflict scholars. Moreover, the KEDS project has spawned a number of similar projects, and this technology has spilled over into a variety of other areas of political science as well.

KEDS and its open-source successor, Text Analysis By Augmented Replacement Instructions (TABARI) program,[1] were originally used to collect information primarily on regional interactions among actors (e.g., the Levant). Although most event data sets code state-to-state interactions, a major breakthrough in the coding of substate actors originated with the Protocol for the Analysis of Nonviolent Direct Action (PANDA) project in the early 1990s. In addition to coding substate actors, PANDA's focus was on acts of nonviolence and low-level contentious politics. That actor system was then incorporated into Integrated Data for Event Analysis (IDEA) system of Bond et al. (1997) that performs global coding. Some of the IDEA data are available publicly whereas much of the data remain proprietary.[2]

Although some of the KEDS project early work focused on examining relations between state and nonstate actors such as the Israelis and Palestinians, a majority of the analyses featured analyses of state-to-state interactions. Such event codes took the form of the

[1]See http://web.ku.edu/keds/index.html for information on the KEDS and TABARI projects. Also see Schrodt (1996, 2006) for the respective codebooks.

[2]Obtain and view the publicly released data on Gary King's dataverse: http://dvn.iq.harvard.edu/dvn/dv/king/faces/study/StudyPage.jsp;jsessionid=5eeb54dd6e1d0ac3ee29a2f69ff6.dvnInstance1?studyId=505.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

World Event Interaction Survey (WEIS) coding scheme and more recently the Conflict and Mediation Event Observations (CAMEO) coding scheme.[3] In 2004, the KEDS group began working on the Political Instability Task Force project, and at that time, it became apparent that there were some problems with existing protocols in terms of coding substate actors and events (Schrodt et al. 2008, 2), which eventually evolved into the CAMEO scheme (Schrodt et al. 2008). PCS builds on this prior event-coding work focusing more closely on substate actors than his work in the past and also focuses on a different region. Specifically, PCS develops a crosscutting hierarchical actor-coding scheme and database detailed down to the individual (e.g., president, Pol Pot) and collects information within and across several countries in South and Southeast Asia.

4 The PCS Innovations, Components, and Workflow

The PCS data differ in various ways from prior event data sets and combine the strengths of many previous data collection efforts. PCS employs automated natural language processing to code the conflictual and cooperative behavior of multiple state, substate, and nonstate actors by subnational location in Cambodia, Indonesia, the Philippines, Burma, Thailand, Bangladesh, Malaysia, Laos, and Vietnam from 1980 to 2006.[4] The coded actions run the gamut from positive and negative statements to bombings to political compromises to armed raids. The result is a record of publicly recorded events and the actors, targets, and locations associated with each event. Employing automated coding methods allows the collection of massive amounts of pertinent information on civil conflict while also eliminating inconsistencies, coder fatigue, and coding time associated with human coding.

PCS uses the TABARI coder to generate domestic political event data. TABARI uses a ''sparse-parsing'' technique to extract the subject, verb, and object from a sentence and determines the appropriate codes using pattern matching on actor and verb dictionaries.[5] The result is a numeric representation of an event in the form of ''someone does something to someone else'' on a certain day. Moreover, to date, we have coded the city or subregional location of events for a sample of countries (e.g., India, Cambodia, Indonesia, and Russia). In sum, the project codes various behaviors in various locations by just about any political actor including individuals (e.g., Pol Pot), groups (e.g., rebels, MILF, etc.), and organizations (e.g., labor unions, military, etc.) for several South and Southeast Asian countries for every day from 1980 to 2006.

4.1 *Constructing the Actor Database*

Machine-coded data are only as good as the dictionaries used to code them. PCS customizes the actor dictionaries for each case and, recently, developed the PCS Actor Database to help standardize the process of dictionary building as well as storing additional information on each of the actors. The database allows for crosscutting hierarchies to be used and implemented by different researchers. The scheme is fundamentally different from other coding schemes as it allows us to group actors based on characteristics and be a part of different hierarchies depending on the question of interest.

For example, it is often the case that an individual can be part of two or three higher level entities. For example, Salamet Hashem is the leader of the MILF, the leader of a separatist

[3]See ''World Event/Interaction Survey (WEIS) Project, 1966–1978,'' ICPSR Study No. 5211 and see Gerner et al. (2002) for information on WEIS and CAMEO, respectively.
[4]Additional countries have been added as funding has increased such as Russia.
[5]TABARI recognizes pronouns and dereferences them. It also recognizes conjunctions and converts passive voice to active voice (Schrodt 1998).

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

group, and the leader of a Muslim group. If we want to study MILF-government interactions, we would code him as part of the MILF. The database also allows us to pool separatist groups and Muslim groups together, and we could easily include him as part of ''separatists'' or as a part of ''Muslim groups.'' Although MILF is a separatist group, it is also a Muslim group. However, not all separatist groups are Muslim, and not all Muslim groups are separatist groups. As such, MILF is part of a crosscutting hierarchy that we can tease out using our database depending on the question or relationships we want to analyse. In addition to categorizing groups into religious, ethnic, and political affiliations, the actor database also contains structural information on each of the groups such as their ideology, structure, information on their recruiting activities, funding sources, etc. Using the information contained in the database, we populate actor dictionaries for each coding project (i.e., each question and/or analysis). We download one news source at a time for a variety of national and local news sources available in Lexis Nexis. After the dictionaries have been developed, we then run the news reports through TABARI using the case-associated actor, verb, and location dictionaries.

In the past, coding frameworks identified the actors a priori as pertinent to each case. More recently, we developed an ''Actor Finder'' software program to search text reports and cull potential relevant actors for each case. We partially generate the lexicon of actors and events by performing part-of-speech tagging using a locally developed tagger and then performing shallow parsing (determination of sentence structure). We are currently experimenting with several approaches, including dependency grammar (where structure is determined by the relation between a word and its dependents) and conventional constituency trees (that divide a sentence into groups of words, e.g., noun and verb phrases), in order to identify potentially relevant nouns that are of interest and then flagging these for human attention, discarding the ones that are not relevant. The remaining nouns then get added to the actor database. In addition to the machine-assisted process, student coders also identify actors within electronic news reports contained in the Lexis-Nexis database.

Given our theoretical priors about micro-level processes and our belief about the number of actors involved in civil conflicts, we attempt to make the actor dictionaries as extensive and disaggregated as possible. Coders collect terms on any major players in society including generic terms (businessmen, religious leaders, dissidents, protestors, or anyone else without a specific name). Once a case is complete, we use the database to generate the dictionary of interest and test its depth by coding events one at a time in TABARI and making necessary changes to our database/dictionaries. We then run TABARI in the ''automated'' mode. Finally, we use a variety of filters to remove duplicate cases and address additional problems.[6]

4.2 *The Actor-Coding Scheme*

At the highest level, PCS groups the codes into general categories (see Table 1) such as government, political parties, dissidents, social actors, and targets. The targets category contains terms that are exclusively targets because they are not able to be the subject of actions, for example a supply depot. Terms can be aggregated along these lines or they can be aggregated at lower levels. Within these general categories, we further disaggregate more specific actor categories. For example, we break up different sectors of the

[6]More detail on filters can be viewed on the KEDS Web site (http://web.ku.edu/keds/index.html) as well as in our online Appendix.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Table 1 Project civil strife general coding scheme

State 0xxx
1– State
2– Administration
3– President/Leading Executive
4– Leading Executive/Head of State
5– Chief of Staff
11–99 High Ranking Executive Officials
100 Legislature
101–199 Legislative Members
Sets of 10s by body
200–299 Judicial Offices
300–399 National Government Positions; Sets of 10
400–499 Sub-National Subdivisions
500 Army
501–519- Army Masses
520–579- Army Elites
580–599- Paramilitary/Secret Police
600–699- Police
700– Other Government Positions
Political parties 1xxx
1– Political Parties
2–499 National Parties
500–999 Local Parties
Done by 10s with 1s digit representing offices in the party
Dissidents 2xxx
1– Dissidents
2–99 Specific Dissident Groups/People
100–199 Generic Rebels/Terrorists
200–999 Case Specific
Social Actors 3xxx, 4xxx
1-Unspecified Social Actor
2–99 Media
100–199 Students and Education
200–299 Classes/Ethnicities (Cross Sections of People)
300–399 Business
400–499 Labor Unions
500–599 Church/Religious Groups
600–4999 Groups and Organizations (by 10s)
Targets 5000–6999
Locations 7000–8999
Miscellaneous 9000–9999

Note. Appendix A contains a sample office coding range for Cambodia.

government such as the executive, the legislature, the judicial branch, the military, and the police and assign each of them specific numerical coding ranges.

Table 1 shows how PCS distinguishes different rebel groups and dissident groups from one another. The coding system uses an 11-digit code, the 3-digit COW country code, and an 8-digit actor code. Terms are coded in PCS based on the actor's role in society. This code matches additional information about the individual or group in the database including the

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

dates of the person's tenure in that office, his or her unique personal ID number (that allows us to track an individual across multiple offices), the office he or she holds, his or her administration affiliation, and his or her organizational status.

The actor code is designed for maximum flexibility in data analysis. The first four digits identify different offices or groups (i.e., 0003 is the code for the president). The next four digits (decimals) indicate the specific person in that office. For instance, for the United States, we would code individuals in the following manner:

0003.0101 George Bush, SR.
0003.0201 William Clinton
0003.0301 George W. Bush

The code 0003 refers to the executive's office, whereas 0101, 0201, and 0301 all refer to specific presidents. To elaborate on the scheme, take the following list of actors and their associated codes:

0003.0201 William Clinton
0003.0301 George W. Bush
0017.0201 Warren Christopher
0017.0202 Madeline Albright
0037.0201 Dee Dee Myers
0017.0301 Colin Powell
0037.0301 Ari Fleischer
0037.0302 Scott McClellan

The first four digits in the codes above refer to the offices: president (0003), secretary of state (0017), and press secretary (0037). The first two decimal places or ''administration'' digits as we like to refer to them designate the different administrations. For example, the code 02 represents the Clinton administration, whereas the 03 code represents the Bush administration. The final two digits differentiate the office holders within the administration. Warren Christopher, for instance, worked in the first (01) Clinton administration (02) as secretary of state (0017), whereas Scott McClellan worked in the second (02) Bush Administration (03) as a press secretary (0037).

The system also allows for key offices or positions inside larger entities to be demarcated. Take for example, the following list of codes:

1510.0000 Democratic Party
1511.0000 Chairman (unknown)
1511.0101 Chairman Harrison
1511.0202 Chairman Uline
1512.0000 Secretary (unknown)
1512.0101 Secretary Slack
1512.0202 Secretary Borna

The example suggests that any code that appears 151x.xxxx is a person within the Democratic Party. 1511 refers to the chairman position and when the person's name is known, the individual Chairmen can be distinguished (1511.0101). Administrations are still preserved such that one can decipher that secretary Borna worked with Chairman Uline.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Finally, PCS also provides a unique ''personal ID'' code for specific people. The personal ID is a four-digit code that uniquely identifies any actor who appears more than once within the data set in different offices or positions. For example, Prince Norodom Sihanouk was head of state, who then was deposed and opposed the government, and then once again became head of state. These personal numbers are useful for tracking individuals who appear on the government and on the dissident side of particular political conflicts. At the same time, they are useful for tracking individuals who hold many different offices or ranks within one organization or government.

Uniquely identifying individual actors and separate groups in the data set provides the ability to study multiple dissident group interactions with the government, interactions among various rebel and ethnic groups, and interactions among actors within a group. To give the reader an idea of the size and complexity of the actor dictionaries, the Cambodia actor dictionary codes over 950 unique individuals and groups. The Indonesia actor dictionary contains 670 unique actors, including more than 50 different dissident organizations and political parties. Both capture well over 200 different distinct positions within the government and military.[7] They also incorporate individuals outside traditional government and dissident groups including political parties, social actors, religious actors, and business groups.

4.3 *Verb Dictionaries*

Our verb dictionary is modified from the KEDS verb dictionaries. Verbs and verb phrases are assigned a category based on the CAMEO coding scheme (Gerner et al. 2002). In analyses, one can use scales or event counts. The KEDS group has developed a CAMEO scale, similar to the Goldstein (1992) scale, which is available on the KEDS Web site.[8] One can also use other mathematical techniques to scale the data such as the Rasch test.[9] Alternatively, one can count events and examine frequencies or alternatively map the CAMEO codes onto another weighting or event category system. For example, Horne, Shellman, and Stewart (2008) mapped the CAMEO codes onto the US government's instruments of national power framework (diplomatic, information, military, and economic).

Because of their labor-intensive nature, human-coding projects made one pass through the text with a single-coding framework. Automated coding, in contrast, allows the use of multiple verb dictionaries or even a dictionary that is developed down the road to code and recode the data. Although we often analyse the conflict or cooperation values of an action, a researcher could conceivably study a variety of tactics, making his or her own unique version of the data by constructing their own verb dictionaries or summing the information in our raw data set in different ways.

4.4 *The Location Dictionaries*

Automated coding can also record the locations of events. Until recently, spatial units relevant to conflict were confined to the state. Buhaug and Gates (2002, 417) argue that ''geographical factors play a critical role in how a civil war is fought and who will prevail.'' The ''location and size of a country'' as well as the location and size of villages, towns, cities, and rebel camps ''affect the design and nature of military strategy'' (Buhaug and Gates

[7]Note that dictionaries often contain multiple phrases/terms for the same actor. As such, 959 unique Cambodian actors are coded using a dictionary with 8844 terms. There are 671 actors and 2007 terms for Indonesia.
[8]See http://web.ku.edu/keds/cameo.dir/CAMEO.SCALE.txt.
[9]See Schrodt (2007) and Horne, Shellman, and Stewart (2008) for examples of scaling event data using item response theory.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Table 2 Number of appearances of the top 10 actors (1980–2004)

Cambodia		*Indonesia*	
Actor name	*No. of times coded*	*Actor name*	*No. of times coded*
Vietnam	9905	Indonesian working class	27 623
Democratic Kampuchea	5867	Indonesian general population	21 695
Royal Cambodian leadership	3768	President Suharto	22 409
Sihanouk (as a dissident)	3500	Indonesian government	22 272
People's Republic of Kampuchea (PRK) government	3902	President Suharto's government	14 201
Guerillas	2375	Minister	10 031
Cambodian Royal Government	2505	Police	7339
Hun Sen	1785	Foreign minister	8188
Sihanouk (as king of Cambodia)	1765	Military	9569
Vietnamese occupation army	2656	ASEAN	8923

2002, 419). Recognizing this fact, we develop three location dictionaries for each case. One contains cities, another contains regions or provinces, and a third records other locations where an event takes place. For example, an event may read, ''US troops face terrorist threats from militants in northern Iraq.'' In this instance, no city or region is given, yet the general location of the conflict is recorded: northern Iraq. Several cities may reside in a region, and several regions may make up an area of the country (e.g., northern Iraq). The scheme allows for one to always aggregate up. In the ''other'' dictionary, we also have codes like ''Thai border.'' Many event data projects simply record an event taking place somewhere in a country. Our location data allow for spatial analysis of conflict and help answer questions concerning contagion and diffusion (Siverson, Martin, and Starr 1991; Buhaug and Gates 2002). For example, Rasler (1996) shows that repression decreased the spatial diffusion of protest in the short run but in the long run repression increased mobilization and protests spread. PCS data can be used for similar analyses.

4.5 *Sources*

Whereas most event data sets (international and intranational) code events from a single news source, we code events from multiple news sources.[10] Davenport and Ball (2002) and Shellman, Reeves, and Stewart (2007) show that media bias influences the scientific inferences we draw from statistical models that analyse data from a single news source. Potentially, language, coverage, style, and characterization by a source can influence the way an event is coded or even if it is coded at all. Schrodt, Simpson, and Gerner (2001, 36) write:

> Reuters and AFP are comparable in terms of the general patterns of events they report. They are not, however, identical sources of information . . .. Reuters provides denser coverage in the Balkans . . .. What seems to be important here is not only that AFP differs in style from Reuters, but that there are regional differences in AFP as well. This suggests that sometimes Reuters is in the right place at the right time, and sometimes AFP.

These findings suggest that source bias deserves more attention, including the possibility of creating multiple source chronologies. PCS codes multiple sources, including local sources, to help average out bias across sources.

[10]For example, early KEDS data come from *Reuters*, whereas later KEDS data come from *Agence France Presse*. WEIS data come from *The New York Times Index*.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Table 3 Dyad event frequency counts (1980–2004)

	All
Cambodia	
Government to Democratic Kampuchea	2634
Democratic Kampuchea to Government	2786
Government to FUNCINPEC	498
FUNCINPEC to government	535
Government to KPNLF	145
KPNLF to government	237
Government to resistance coalition	1208
Resistance coalition to government	1818
Indonesia	
Government to Jemaah Islamiyah	440
Jemaah Islamiyah to government	159
Government toward Free Aceh Movement	354
Free Aceh Movement toward government	269
Government toward Free Papua Movement	103
Free Papua Movement toward government	121
Government toward East Timor Rebels	306
East Timor Rebels toward government	357

5 The PCS Data Sets

This section describes data we collected for Cambodia and Indonesia 1980–2004. Table 2 reports the frequencies for the top 10 actors who appear in the PCS Cambodia and Indonesia data sets. Vietnam is one of the most frequent actors in the Cambodia data set because the Vietnamese government controlled Cambodia for quite some time. The rebel group, Democratic Kampuchea, ranks second, followed by Sihanouk as a dissident and later as king. Different governments and leaders like Hun Sen as well as the Vietnamese Occupation Army round out the top 10. The frequency of events that involve Sihanouk as a dissident and as a rebel shows how PCS is able to track different individuals over time across different offices and roles. In terms of Indonesia, the working class and population show up frequently. The government, Suharto, the police, and the military are also quite frequent.

Typically, scholars study interactions between actors and to do so often create directed dyadic measures of conflict and cooperation. Table 3 shows the number of times particular directed dyads are involved in events. Government-Democratic Kampuchea interactions dominate the directed dyadic events in Cambodia, whereas in Indonesia interactions between the government and Jemaah Islamiyah, the Free Aceh Movement, and the East Timor separatists are more balanced across the major groups. For an example of multiactor models of conflict using the PCS data, see Shellman, Hatfield, and Mills (n.d.).

Aside from examining different groups' interactions with a unitary government, the PCS data allow scholars to examine relationships between groups and specific government leaders, branches of government, and bureaucratic agencies within governments. Table 4 illustrates the different disaggregated government dyads that could be studied using the PCS data. An X refers to the specific actor or government branch heading the right-hand side columns. For example, the data show that the executive branch takes 2516 actions toward all dissident groups during the 1980–2004 period of time. Table 4 also shows that Democratic Kampuchea takes 774 actions toward the military during the time period. All

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Table 4 Dyad frequency counts with a disaggregated government (1980–2004)

	Disaggregation of government (X)				
	Executive	*Judicial*	*Legislative*	*Military*	*Police*
Cambodia					
X to all dissident	2516	47	182	1106	312
All dissident to X	1385	22	89	1824	213
X to Democratic Kampuchea	970	15	56	354	22
Democratic Kampuchea to X	362	13	16	774	22
Indonesia					
X to all dissident	14659	689	1017	2583	3352
All dissident to X	15104	442	1044	4425	2504
X to Jemaah Islamiyah	116	43	7	35	131
Jemaah Islamiyah to X	42	11	2	7	30
East Timor Rebels toward X	154	5	9	53	13
X toward East Timor Rebels	168	3	11	72	12

in all, the table reveals that different relationships among various posts and branches in the government and other political actors can be analysed using these data.

Finally, Fig. 1 plots some preliminary data that PCS has collected on the locations of conflict events in Cambodia. The two most populated regions in terms of conflict events include Phnom Penh and Pailin. Phnom Penh is the capital of Cambodia and the central location of the government, whereas Pailin was a major stronghold and resource center, rich in gem stones, for the Khmer Rouge. The preliminary spatial plot supports other findings in the literature. For example, conflict takes place in the resource-rich areas and in the mountains (Pailin and Kampot). Moreover, the abundance of conflict events seems to be greater near the Vietnam border (Prey Veaeng, Sbaay Rieng, Kampong Chaam, and Kracheh). In sum, the PCS data can be used to study multiple-actor and group interactions over disaggregated time and space.

6 Conclusion: New Avenues for Analysis and Event Coding

The goal of PCS is to improve the quality of event data and allow for new theories of civil conflict to be tested by collecting highly disaggregated event data from numerous sources that yield information on the behavior of multiple substate actors in small temporal (i.e., days) and spatial (cities, regions, etc.) units.

6.1 *New Avenues for Analysis*

Most traditional models of civil conflict focus on the national attributes of a country that give rise to intrastate conflict. Other more recent studies focus on dyadic interactions between a dissident group and a government (Moore 1998, 2000; Shellman 2006a, 2006b; Heger and Salehyan 2007). However, we know that many intrastate conflicts are fought among multiple groups, and the state and some countries may be experiencing many different conflicts at the same time (e.g., India). Moreover, most intrastate conflicts do not span the entire country and often entail escalations and de-escalations in conflict and cooperation. Various groups use different tactics at different times, and such conflicts often cannot be summed up in a single annual summary measure. As such, we need to develop theoretical and empirical models to explain, analyse, and predict the ebb and flow of

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

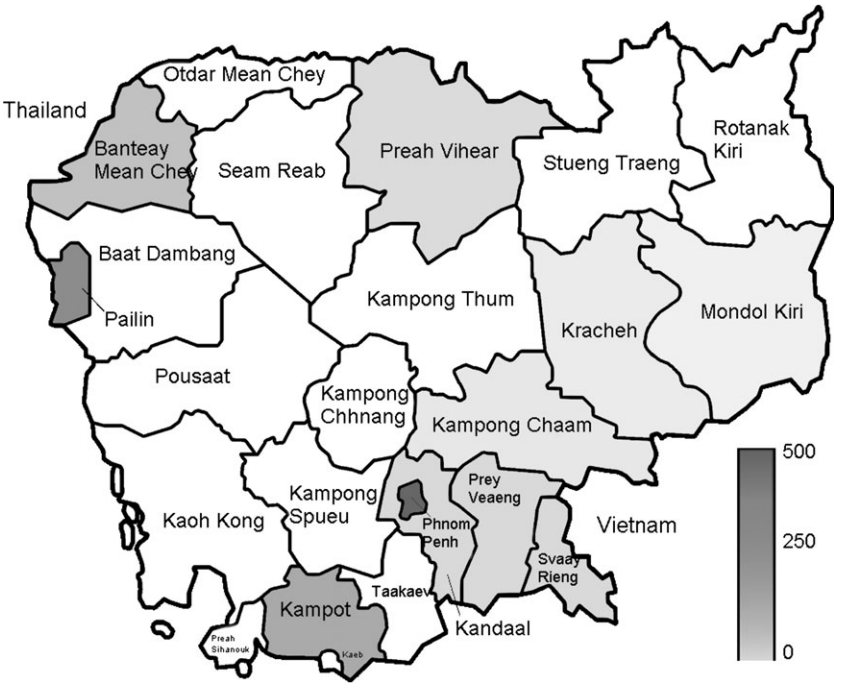


Fig. 1 Regional disaggregation of conflict events in Cambodia, 1980–2004.

conflict and cooperation among multiple groups and the governments they challenge over a finer temporal and spatial domain.

The PCS data sets respond to these challenges. Innovations in technology and computing power have enabled disaggregated data to be collected quickly and accurately from open-source text. Moreover, innovations in actor-coding schemes have allowed us to collect information on the behavior of myriad actors, who often represent different offices or sides of a conflict altogether, across both time and space. The data will hopefully generate interest into other research questions and stimulate thought on future models that could be constructed and tested.

6.2 *New Avenues for Automated Event Coding*

This study demonstrates the current and future possibilities of the use of automated natural language processing to perform content analysis in the field of intrastate conflict. Machine coding allows us to collect disaggregated events across multiple actors, space, and time at an unprecedented speed. King and Lowe (2003, 636) show that the performance of machine coders to human coders is virtually identical. Time and money can be saved while at the same time advances are made in refining the type of data we collect using automated natural language processing methods.

The future of automated event coding is bright. Most existing automated event coders focus on the English language. The next step is to delve into foreign language sources. Arguably, there is a wealth of information in such sources that most likely differs from what is reported in English sources. Moreover, the sentence is the unit of analysis for most current coders. This poses several problems including but not limited to coding duplicate events and coding information in a current news report about events in the past. The future

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

of event coding can alter the unit of analysis to the report and perhaps in the not too distant future the event itself. Being able to code several reports about a single event yields some complimentary and some redundant information. As it stands now, the redundancies are often unfiltered, and the complementary information is treated as distinct and discrete.

Although TABARI and other coders can co-reference pronouns within the same sentence, they cannot co-reference nouns within the report. For example, a report may lead off with ''the MILF attacked a military base'' and later note that ''the rebels kidnapped a military official'' as part of the attack. Although TABARI would code each of these events, it would not be able to decipher that MILF was responsible for the kidnapping. Future coders can tackle these intra-report and noun co-referencing issues as well as attempting to make the event the unit of analysis. Moreover, with advances in computational linguistics, we can begin to extract hidden meaning from texts such as sentiment and/or signaling behavior and/or study the art of diplomacy as it plays out across actors who are party to a conflict. There is more theorizing to be done, more data to be collected, and certainly more models to be run but future developments in natural language processing and artificial intelligence are likely to aid our search for the elusive truth.

Funding

National Science Foundation (SES-0516545 & SES-0619997).

References

Bond, Doug, J. Craig Jenkins, Charles L. Taylor, and Kurt Schock. 1997. Mapping mass political conflict and civil society: The automated development of event data. *Journal of Conflict Resolution* 41(4):553–79.

Buhaug, Halvard, and Scott Gates. 2002. The geography of civil war. *Journal of Peace Research* 39(4):417–33.

Collier, Paul, and Anke Hoeffler. 2004. Greed and grievance in civil wars. *Oxford Economic Papers* 56:663–595.

Davenport, Christian. 1995. Multi-dimensional threat perception and state repression: An inquiry into why states apply negative sanctions. *American Journal of Political Science* 38(3):683–713.

Davenport, Christian, and Patrick Ball. 2002. Views to a kill: exploring the implications of source selection in the case of Guatemalan State Terror, 1977–1995. *Journal of Conflict Resolution* 46(3):427–50.

Fearon, James, and David Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1):75–90.

Gerner, Deborah J., Philip A. Schrodt, Ömür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (CAMEO): A new event data framework for a post cold war world. Presented at the annual meeting of the American Political Science Association, 29 August–1 September 2002. http://web.ku.edu/keds/papers.dir/gerner02.pdf (accessed October 8, 2008).

Goldstein, Joshua S. 1992. A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution* 36(2):369–85.

Goldstein, Joshua S., and J. C. Pevehouse. 1997. Reciprocity, bullying, and international cooperation: Time-series analysis of the Bosnia conflict. *American Political Science Review* 91(3):515–30.

Heger, Lindsay, and Idean Salehyan. 2007. Ruthless rulers: coalition size and the severity of civil conflict. *International Studies Quarterly* 51(2):385–403.

Hegre, Håvard, Tanja Ellingsen, Scott Gates, and Nils Petter Gleditsch. 2001. Toward a democratic civil peace? Democracy, political change, and civil war 1816-1992. *American Political Science Review* 95(1):16–33.

Horne, Cale D., Stephen M. Shellman, and Brandon M. Stewart. 2008. *Nickel and DIMEing the adversary: Does it work or PMESII them off?* Presented at the annual meeting of the International Studies Association, San Francisco, CA, 26–29 March.

King, Gary, and Will Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization* 57(3):617–42.

Lichbach, Mark. 1987. Deterrence or escalation? The puzzle of aggregate studies of repression and dissent. *Journal of Conflict Resolution* 31(2):266–97.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press

Moore, Will H. 1998. Repression and dissent: Substitution, context, and timing. *American Journal of Political Science* 42(3):851–73.

Moore, Will H. 2000. The repression of dissent: A substitution model of government coercion. *Journal of Conflict Resolution* 44(1):107–27.

Moore, Will H. 2006. A problem with peace science: The dark side of COW. Working paper. http://mailer.fsu.edu/~whmoore/garnet-whmoore/research/DarkSideofCOW.pdf (accessed October 8, 2008).

Rasler, Karen. 1996. Concessions, repression, and political protest in the Iranian revolution. *American Sociological Review* 61(1):132–52.

Rød, Jan Ketil, and Halvard Buhaug. 2007. Civil wars: Prospects and problems with the use of local indicators. Paper presented at ScanGIS'2007 Conference. http://www.scangis.org/scangis2007/papers/e7_rod.pdf (accessed October 8, 2008).

Sambanis, Nicolas. 2001. Do ethnic and nonethnic civil wars have the same causes? *Journal of Conflict Resolution* 45(4):259–282

Schrodt, Philip A. 1998. KEDS: Kansas event data system manual. http://web.ku.edu/keds/software.dir/keds.html (accessed October 8, 2008).

Schrodt, Philip A. 2006. TABARI manual, version 0.5. http://web.ku.edu/keds/tabari.dir/tabari.manual.060228.pdf (accessed October 8, 2008).

Schrodt, Philip A., Shannon G. Davis, and Judith L. Weddle. 1994. Political science: KEDS—A program for the machine coding of event data. *Social Science Computer Review* 12(4):561–87.

Schrodt, Philip A., and Deborah J. Gerner. 1994. Validity assessment of a machine-coded event data set for the Middle East, 1982–1992. *American Journal of Political Science* 38(3):825–54.

Schrodt, Philip A., Erin M. Simpson, and Deborah J. Gerner. 2001. *Monitoring conflict using automated coding of newswire sources: A comparison of five geographical regions.* Paper presented at the PRIO/Uppsala University/DECRG High-Level Scientific Conference on Identifying Wars: Systematic Conflict Research and Its Utility in Conflict Resolution and Prevention, Uppsala, Sweden 8–9 June 2001.

Schrodt, Philip A. 2008. TABARI manual, version 0.6.3B7. Lawrence, KS: University of Kansas, Lawrence. http://web.ku.edu/keds/tabari.dir/tabari.manual.0.6.3b7.pdf (accessed October 8, 2008).

Schrodt, Philip A. 2007. Inductive event data scaling using item response theory. Paper presented at the annual meeting of the International Studies Association 48th Annual Convention. http://www.allacademic.com/meta/p179617_index.html (accessed October 8, 2008).

Schrodt, Philip A., Omur Yilmaz, Deborah J. Gerner, and Dennis Hermrick. 2008. Coding sub-state actors using the CAMEO (conflict and mediation event observations) actor coding framework. Version 1.0B1. Presented at the International Studies Association meeting, San Francisco, 26–29 March. http://web.ku.edu/keds/papers.dir/ISA08.pdf (accessed October 8, 2008).

Shellman, Stephen M. 2006a. Leaders & their motivations: Explaining government-dissident conflict-cooperation processes. *Conflict Management & Peace Science* 23(1):73–90.

Shellman, Stephen M. 2006b. Process matters: Conflict & cooperation in sequential government-dissident interactions. *Security Studies* 15(4):563–99.

Shellman, Stephen M., Clare J. Hatfield, and Maggie J. Mills. Disaggregating actors in intranational conflict. *Journal of Peace Research.* Forthcoming.

Shellman, Stephen M., Andrew M. Reeves, and Brandon M. Stewart. 2007. Fair & balanced or fit to print: The effects of source bias on event data analysis. http://web.ku.edu/keds/papers.dir/forecasting.html (accessed October 8, 2008).

Siverson, Randolf Martin, and Harvey Starr. 1991. *The diffusion of war: A study of opportunity and willingness*. Ann Arbor, MI: University of Michigan.

Stephen M. Shellman. 2004. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis* 12:97–104.

Thyne, Clayton. 2006. Cheap signals with costly consequences: The effect of interstate relations on civil war, 1948–1992. *Journal of Conflict Resolution* 50(6):937–961.

Tilly, Charles. 1985. *Models and realities of popular collective action* Working paper no. 10, Center for Studies of Social Change. New York: New School for Social Research.

https://doi.org/10.1093/pan/mpn008 Published online by Cambridge University Press